

# **FORECASTING THE FUTURE WATER HEIGHT OF A RESERVOIR WITH TEMPORAL DATA MINING**

Mohamed Saudulla Aabid Rushdi

168265J

Thesis submitted in partial fulfillment of the requirements for the degree  
Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

June 2018

## DECLARATION

I declare that this is my own work and this thesis does not incorporate without acknowledgement, any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works.

.....  
Mohamed Saudulla Aabid Rushdi  
Date

The above candidate has carried out research for the Masters thesis under my supervision.

.....  
Dr. Amal Shehan Perera  
Date

## **ABSTRACT**

Throughout last few decades, flooding has been one of the most expensive natural disasters, which has had an increased impact on human casualties, property damage and rehabilitation. Thus, there is a crucial need in taking necessary actions to avoid or minimize the impact caused by floods to human lives as well as to the stability of the economy of the country. As to the most recent devastating flood experience faced in Sri Lanka in 2016, one of the root causes identified was the unpredictable decision making strategy, which was used to control and manage excess water in reservoirs due to heavy rain. Moreover, reservoirs are a key water storage source in water management, where they are utilized by various sectors for different purposes. Therefore, there is an essential need in taking the best decision in releasing water from reservoirs to minimize the damage caused by the floods and to optimize the utilization of water as a scarce resource.

Many researches have been carried out on the field of data science based on collected data from rivers, reservoirs & tanks to support decision making in water management. Those are mainly focused on classifying water release rules and ranges of a reservoir or tank. Research on forecasting the future water height of reservoirs, with both rainfall and uncertain water inflow due to human intervention, are very limited.

Hence, this research focuses on predicting the future water height of a reservoir, when the water inflow is uncertain and the reservoir receives a significant amount of rainfall. This would allow to minimize the risk of deadly floods by opening the sluice gates in time and to manage the water in an optimum manner for irrigation purposes. This research proposes the most effective set of features to forecast the water height of a reservoir on next three days. Furthermore, it presents the comparison of the performance of different regression models and the effectiveness of applying clustering techniques on top of the regression models. The result obtained from this research demonstrates that, K-Medoids clustering with feed forward artificial neural network model has the best performance in forecasting the reservoir water height when there is a significant amount of rainfall and the water inflow is uncertain.

## ACKNOWLEDGEMENT

There is a tremendous amount of hard work and commitment behind the success of this research and I take this opportunity to convey my heartiest gratitude towards all the individuals who were a part of it. It is indeed my privilege to thank each one of them who gave me their valuable support, constant and necessary guidance while giving me the right level of motivation towards reaching this research's ultimate goal.

Firstly, I would like to express my heartfelt gratitude Dr. Amal Shehan Perera, Department of Computer Science and Engineering; as my project supervisor for his remarkable support, supervision and the inspiring advices given throughout, to make this research a success.

While my sincere thanks also go to Dr. Malaka Walpola, Dr. Charith Chitraranjan, Dr. Surangika Ranathunga and Dr. Indika Perera for their valuable feedback, encouragement and advices through time to time evaluations and providing me with the necessary advice and guidance in initiating and carrying out this research.

I also take this opportunity to thank Mr. Sanjaya Rathnayaka, and Mahaweli Authority for providing me the required data and helping me with the domain knowledge. I also should be grateful to darksky.net and timeanddate.com websites, from where I have collected some of the necessary data for the simulation.

Finally, I would like thank to each individual; especially my parents for their love & motivation and my loving wife for her support towards making this research a success during this entire period of time. I would also like to express my thanks to all my colleagues at MSc 16 batch and at Synopsys Lanka (PVT) Ltd.

## TABLE OF CONTENTS

DECLARATION .....	i
ABSTRACT .....	ii
ACKNOWLEDGEMENT .....	iii
TABLE OF CONTENTS .....	iv
LIST OF FIGURES .....	vi
LIST OF TABLES .....	ix
LIST OF ABBREVIATIONS .....	x
1 INTRODUCTION .....	1
1.1 Water Management.....	1
1.2 Water Reservoirs and Tanks.....	2
1.3 Cascaded Tank System.....	4
1.4 Knowledge Discovery in Data.....	6
1.5 The Research Problem.....	7
1.6 The Motivation .....	8
1.7 Reservoir in Study – Maduru Oya Reservoir .....	9
2 LITERATURE REVIEW .....	12
2.1 Statistical Approach.....	12
2.2 Temporal Data Mining .....	14
2.2.1 Artificial Neural Network (ANN) models .....	14
2.2.2 Other prediction models .....	16
2.2.3 Reservoir water height forecasting models .....	16
2.3 Spatio–temporal data mining.....	19
2.4 Water disaster management with data mining.....	19
2.5 Predicting the uncertainty of rainfall .....	21
3 METHODOLOGY .....	22
3.1 Data collection.....	22
3.2 Data preprocessing .....	23
3.3 Building the machine learning model.....	25
3.3.1 Statistics of the data.....	25
3.3.2 Developing the model .....	26

3.4	Evaluating the features and models .....	28
4	Solution Architecture .....	30
4.1	User input .....	30
4.2	Output .....	30
4.3	Training flow of the system.....	30
4.4	Prediction flow of the system .....	31
5	Results Analysis And Discussion .....	35
5.1	Feature selection .....	35
5.2	Clustering .....	42
5.2.1	Number of clusters .....	43
5.3	Regression model .....	48
5.3.1	Linear regression model .....	48
5.3.2	MLP regression model .....	57
5.4	Results comparison.....	67
6	Conclusion and future work.....	70
6.1	Conclusion .....	70
6.2	Future work .....	71
	REFERENCES.....	73

## LIST OF FIGURES

	Page	
Figure 1.1	Components and their relative positions of a small tank in Sri Lanka	3
Figure 1.2	Schematic of a tank and the water balance components	3
Figure 1.3	Cascaded irrigation tanks in Anuradhapura, Sri Lanka	5
Figure 1.4	Location of Maduru Oya Reservoir	9
Figure 1.5	Ancient Maduru Oya Sluice of 1 <sup>st</sup> century BC	10
Figure 1.6	Details of Maduru Oya Reservoir Project	11
Figure 2.1	Thirappane cascade system model	12
Figure 3.1	Daily rainfall in year 2009	24
Figure 3.2	Daily rainfall in year 2013	24
Figure 3.3	Human errors in water head height	25
Figure 3.4	Possible approaches	28
Figure 4.1	Overview of the system	31
Figure 4.2	Training the machine learning models	32
Figure 4.3	Prediction flow of the machine learning model	33
Figure 5.1	Correlation coefficients of the features to predict the water height on day 1	37
Figure 5.2	Variation of length of day time within a year	37
Figure 5.3	Correlation coefficients of the features to predict the water height on day 2	40
Figure 5.4	Correlation coefficients of the features to predict the water height on day 3	42
Figure 5.5	Variation of average RMSE value of linear regression model, with various clustering techniques	43
Figure 5.6	Variation of average RMSE of predicted water height of linear regression model on day 1 with number of clusters on cross validation data set	46
Figure 5.7	Variation of RMSE of predicted water height of linear regression model on day 1 with number of clusters on test data set	46
Figure 5.8	Variation of average RMSE of predicted water height of linear regression model on day 2 with number of clusters on cross validation data set	46
Figure 5.9	Variation of RMSE of predicted water height of linear regression model on day 2 with number of clusters on test data set	46

Figure 5.10	Variation of average RMSE of predicted water height of linear regression model on day 3 with number of clusters on cross validation data set	46
Figure 5.11	Variation of RMSE of predicted water height of linear regression model on day 3 with number of clusters on test data set	46
Figure 5.12	Variation of average RMSE of predicted water height of MLP regression model on day 1 with number of clusters on cross validation data set	47
Figure 5.13	Variation of RMSE of predicted water height of MLP regression model on day 1 with number of clusters on test data set	47
Figure 5.14	Variation of average RMSE of predicted water height of MLP regression model on day 2 with number of clusters on cross validation data set	47
Figure 5.15	Variation of RMSE of predicted water height of MLP regression model on day 2 with number of clusters on test data set	47
Figure 5.16	Variation of average RMSE of predicted water height of MLP regression model on day 3 with number of clusters on cross validation data set	48
Figure 5.17	Variation of RMSE of predicted water height of MLP regression model on day 3 with number of clusters on test data set	48
Figure 5.18	Variation of actual and predicted water height of cross-validation data set on day 1, with linear regression model	51
Figure 5.19	Variation of actual and predicted water height of test data set on day 1, with linear regression model	52
Figure 5.20	Variation of actual and predicted water height of cross-validation data set on day 2, with linear regression model	53
Figure 5.21	Variation of actual and predicted water height of test data set on day 2, with linear regression model	54
Figure 5.22	Variation of actual and predicted water height of cross-validation data set on day 3, with linear regression model	55
Figure 5.23	Variation of actual and predicted water height of test data set on day 3, with linear regression model	56
Figure 5.24	Variation of predicted water height against actual water height on day 1 of cross-validation data set, with MLP regression model	58
Figure 5.25	Variation of predicted water height against actual water height on day 1 of test data set, with MLP regression model	58



Figure 5.26	Variation of predicted water height against actual water height on day 2 of cross-validation data set, with MLP regression model	59
Figure 5.27	Variation of predicted water height against actual water height on day 2 of test data set, with MLP regression model	59
Figure 5.28	Variation of predicted water height against actual water height on day 3 of cross-validation data set, with MLP regression model	59
Figure 5.29	Variation of predicted water height against actual water height on day 3 of test data set, with MLP regression model	59
Figure 5.30	Variation of actual and predicted water height of cross-validation data set on day 1, with MLP regression model	61
Figure 5.31	Variation of actual and predicted water height of test data set on day 1, with MLP regression model	62
Figure 5.32	Variation of actual and predicted water height of cross-validation data set on day 2, with MLP regression model	63
Figure 5.33	Variation of actual and predicted water height of test data set on day 2, with MLP regression model	64
Figure 5.34	Variation of actual and predicted water height of cross-validation data set on day 3, with MLP regression model	65
Figure 5.35	Variation of actual and predicted water height of test data set on day 3, with MLP regression model	66

## LIST OF TABLES

		Page
Table 3.1	Features in the dataset	22
Table 3.2	Statistical parameters of the training data set	26
Table 3.3	Statistical parameters of the test data set	26
Table 3.4	Necessity of sliding window technique	27
Table 5.1	Pearson correlation coefficient of different features with the height of water on day 1	36
Table 5.2	Pearson correlation coefficient of different features with the height of water on day 2	39
Table 5.3	Pearson correlation coefficient of different features with the height of water on day 3	41
Table 5.4	Variation of RMSE value of predicted water height from linear regression model, with the number of clusters	44
Table 5.5	Variation of RMSE value of predicted water height from MLP regression model, with the number of clusters	45
Table 5.6	Performance of the linear regression model on cross-validation and test data sets	49
Table 5.7	Performance of the linear regression model on cross-validation and test data sets, when data sets are clustered	49
Table 5.8	Performance of the MLP regression model on cross-validation and test data sets	57
Table 5.9	Performance of the MLP regression model on cross-validation and test data sets, when data sets are clustered	58
Table 5.10	Performance of the MLP regression model on cross-validation and test data sets with sigmoid activation function	60
Table 5.11	Comparison of water height prediction models	69

## LIST OF ABBREVIATIONS

<b>Abbreviation</b>	<b>Description</b>
ANFIS	Adaptive Neuro-Fuzzy Inference System
ANN	Artificial Neural Network
AR	Auto-Regressive
ARMA	Auto-Regressive Moving Average
CHRS	Center for Hydrometeorology and Remote Sensing
Day 0	The day from which the future water height is forecasted
Day 1	Following day to a given date (day 0), on which the water height is forecasted
Day 2	Second consecutive day from day 0, on which the water height is forecasted
Day 3	Third consecutive date from day 0, on which the water height is forecasted
GEP	Gene Expression Programming
GIS	Geographic Information System
IWMI	International Water Management Institute
MLP	Multilayer Perceptron
MSE	Mean Squared Error
NN	Neural Network
PCA	Principal Component Analysis
RF-SVR	Reduced Feature - Support Vector Regression
RMSE	Root Mean Square Error
SPI	Standardized precipitation index
SVM	Support Vector Machine
SVR	Support Vector Regression
TDANN	Time Delay Artificial Neural Network